DOCUMENT RESUME

ED 091 445 TM 003 663

AUTHOR	Finkel, A.; Norman, G. R.
TITLE	The Validity of Direct-Observation in Assessment of Clinical Skills.
PUB DATE	[73]
NOTE	6p.
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Clinical Experience; Comparative Analysis; Evaluation Techniques; Formative Evaluation; Medical Students; *Observation; Reliability; *Skills; Supervisors; *Validity
IDENTIFIERS	Canada

ABSTRACT

Two modes of evaluation are compared: the summary evaluation by supervisors performed at six-month intervals, and the technique of direct observation of a clinical encounter through one-way glass. The sample consists of 17 residents in pediatrics who were evaluated, using both methods, over an eight-month interval. The analysis of data indicates that the reliability of the direct observation technique is acceptable, in contrast to the low reliability of the supervisor's assessment. A positive correlation exists between evaluations obtained from each method, suggesting that the two methods are measuring the same behaviors, but the results are not significant, probably because of the low reliability of the supervisor's assessment. Finally, both methods showed the expected change with educational level, with the direct observation scores displaying a change of two to three times the supervisor's assessments. The results indicate that the method of direct observation is a more reliable and valid assessment technique than the assessment by supervisors. The implications of this conclusion are discussed. (Author/BB)

THE VALIDITY OF DIRECT-OBSERVATION IN ASSESSMENT OF CLINICAL SKILLS

A. Finkel, M.D., and G. R. Norman, Ph.D.

Introduction:

Recent recognition has been given to the inadequacy of present testing instruments in the assessment of clinical skills. Ample evidence has accumulated that the traditional information-oriented examination correlates poorly with subsequent clinical performance[1]. This evidence has led to innovations in the certification examination[2,3], use of formative evaluations as a component of the certification process[4], and directives for investigation of new evaluation techniques at the national level[5].

Since the majority of these evaluation techniques involve some degree of simulation of the physician-patient encounter, ranging in fidelity from the use of one-way glass to observe a workup of the real patient, to the paper-and-pencil format of the Patient Management Problem[6], it is essential to examine both the internal reliability of the method, and its external validity. One problem in establishing the validity of any evaluation of clinical skills is the absence of any objective measure of clinical competence, and the validity must generally be inferred from indirect analyses.

In the present paper, we focus on the direct-observation of the clinical workup, using either real or simulated patients. The reliability of the technique has been established[7], and preliminary data suggest a positive correlation with a similar assessment by clinical supervisors. In the present work, the independent assessment by clinical supervisors, an evaluation mode which has gained widespread acceptance[4], will be examined in greater detail, and the relative validity of the two methods inferred from an analysis of the reliability of each method, examination of change in evaluations with educational level, and a correlation between methods.

Materials and Methods:

a) Residents

Of the 15 Pediatric residents evaluated, eight were first year and seven were second year residents. Their medical backgrounds varied greatly; most had graduated from medical schools in their country of origin and had been in Canada for varying periods of time. The residents spent three month rotations on nursery, in-patient or ambulatory services in the Pediatric program.

b) Evaluators

Evaluators were seven general Pediatricians in consulting practice, who were heavily involved in patient care and serve as attending physicians on the various services in the residency program.

Reprints: G. R. Norman, Ph.D
Faculty of Medicine, McMaster University
1200 Main St. West
Hamilton, Ontario L8S 4J9
Canada

c) Long Term Assessment

An evaluation form was sent to two or more physicians who had had the most contact with the resident during his rotation in a particular service. The physicians were generally those attending on that service or those admitting many patients to the service during the resident's rotation. An accompanying letter was sent to these physicians asking them to complete the forms if they felt they had had sufficient contact with the resident to evaluate his work.

The scores and comments of each evaluator were summarized and feedback occurred through the small tutorial groups in which each resident participated. Scores were derived by assigning numerical values to the categories on the form. These evaluations were completed in October 1972 and February 1973 requesting on each occasion reports covering the previous three months rotation.

d) Single Encounter

In this form of evaluation, the resident was observed doing a history and physical examination on a patient, the observers watching from behind a one-way viewing screen equipped with audio facilities. Teams of two evaluators who were general consulting pediatricians were selected. The patients were those of one of the two evaluators. The patients were fully informed prior to being seen of the fact that they would be observed and their consent obtained. The patients selected for junior residents were generally single problems (obesity, enuresis, abdominal pain) while those selected for senior residents were more complex (syndromes associated with mental retardation and behavioural disorders, or complex multi-system diseases).

On two occasions because of patient cancellations it was necessary to use a programmed patient. An infant or toddler from the adjoining pediatric ward was picked as the patient. A nurse from the ward was given a prepared history which coincided with the child's clinical status. The nurse was coached a few days prior to the evaluations as to how to perform as a programmed mother. During evaluations involving this patient neither evaluators or residents had prior knowledge nor suspected afterwards that the nurse was not the child's real mother.

Instructions to evaluators were as follows:

Evaluators were to compare the resident's performance to that of an expert pediatrician. They were asked to become familiar with the evaluation form which outlined specific areas in which the resident's performance was to be judged. They were given brief summaries of the patients' problems which listed pertinent, negative and positive features of the history and physical and which included the suggested plan of management. By using this summary and by taking brief notes while observing the resident the evaluator could compare the history and physical obtained by the resident to the summarized findings and could see errors of omission and technique. During observation, the evaluator could score all parts of the evaluation form except for problem formulation and plan of management, both of which were scored after the resident presented this information to the evaluators in the feed-back session following the observation.

The explanations given to the residents prior to this form of evaluation stressed that the evaluation was to be viewed as an exercise rather than an examination. The residents were told that the patients they would see would

generally be those of a pediatrician who would be evaluating them. The patient would be returning for follow-up of a particular problem and the resident's goal in seeing the patient was to determine the nature of the chronic problem as well as the current clinical status referrable to that problem. It was explained that the resident would be observed during history and physical examination and would then be expected to formulate a plan of investigation and management appropriate for the patient's current problems. The initial part of the feed-back session with the evaluators would be the resident's presentation of the patient's problems as he saw them, his plans for investigation and management of those problems.

These evaluations took place in October 1972 and March 1973. Two days were scheduled for the evaluation of 15 residents on each occasion. Patients were given consecutive one-hour appointments. Evaluator teams usually worked for one-half day. The evaluation of one resident took place in one hour. The resident was allowed forty minutes with the parent and child during which he was monitored by the evaluators in the viewing room. In the next twenty minutes, the resident met with the evaluators and presented the patient to the evaluators. This was followed by a discussion of the resident's performance with him by his evaluators.

Analysis of Data:

Data analysis was directed to an investigation of firstly, the internal consistency or reliability of each category for the single encounter (SE) and long-term (LT) evaluations, and secondly, by examining the correlations between SE and LT assessments, and the change in evaluations over the eight-month interval, the validity of each method.

As an initial step, distributions of raw scores accumulated over all categories and all evaluators were plotted as shown in Figure I. From the figure, it is evident that the SE scores are distributed broadly over the range of possible values, with a calculated standard deviation of 0.98. By contrast, the LT estimates follow a much narrower distribution, with 85% of all scores falling in the range 3-4, and a standard deviation of 0.49. Secondly, 3% of SE evaluations fell in the "not applicable" category versus 9% of LT estimates. These results provide a measure of the ability of the instrument to discriminate levels of performance, and indicate that the SE evaluations have greater discrimination.

The raw scores were then utilized in a calculation of reliability, using the method of split-halves, and the Spearman-Brown formula[8]. Reliability coefficients for each category are shown in Table I.

Considering first the SE evaluations, thirteen of the eighteen categories had reliabilities greater than 0.45. Two categories, Investigations and Treatment, had reliabilities of about 0.3, and three categories, Problem Orientation of History, Priority of Problems, and Disposition, had reliabilities in the range 0 to 0.1. It is evident that difficulties were present in assessing problem formulation and management, a result at variance with previous analyses using this form[7]. The difficulties may be due in part to the assessment of problem formulation in discussion with the resident rather than from a written record, the method formerly utilized. The low reliability of the history category is difficult to rationalize, as other similar categories had high reliability. The average of all reliabilities was 0.50.

Turning to the LT estimates, [illegible] of [illegible] ranged from [illegible] to [illegible], with an average value of [illegible]. The results are not surprising in view of the narrow distribution of raw scores. Nevertheless, the low reliability of the data raises serious questions regarding the utility of this approach to performance evaluation.

Analysis of Validity:

In analysing the validity of any evaluation of clinical performance, the basic question to be answered is "How well does the evaluation instrument reflect the habitual clinical performance of the subject physician?" The question is particularly cogent when applied to the certification process, and one ramification has been the use of ongoing evaluation, similar to the long-term assessment of the present work, as a component of the certification grading by the Royal College of Physicians and Surgeons of Canada[4]. However the analysis of reliability in the preceding section would indicate that this method is in itself sufficiently unreliable to raise questions about the external validity of such evaluations.

Since, at the present time, there is no independent, reliable measure of clinical skills with which the SE and LT evaluations could be compared, the analysis of validity was approached indirectly, by first correlating evaluations from each method to ascertain if the two methods were assessing the same characteristics, and by examining the change in evaluations from September to March (construct validity).

Since the categories assessed in each method were not identical, a first step was to group categories, and average scores, in such a way as to develop common characteristics.

Correlation coefficients for the seven grouped categories are shown in Table II. Six coefficients are positive,but none reach significance at the .05 level. Two conclusions may be drawn from these results; that the different methods are assessing different characteristics, or that the unreliability of the LT estimates precludes any meaningful comparison with other measures.

Analysis of the change in evaluations from September to March is shown in Table II. It will be noted, that although all changes are in the positive direction, change in the SE estimates is approximately twice the observed change in LT data. An interesting observation is that the category which least changes in the SE assessments, (G-Patient Interaction), is that which shows the greatest change in the LT estimates. Since the clinical supervisor rarely observes the resident in a one-to-one relationship with patients, it is postulated that the large change in the LT estimates is a reflection of the supervisors *own* greater familiarity with the resident. If this category is removed from the average, the average change in SE estimates is 0.300, compared with 0.109 for the LT estimates.

Discussion of Results:

The analysis of reliability indicates that the SE method results in subjective evaluations with a fair degree of reliability. Certain areas, particularly problem formulation and management were inadequate, and may be improved by assessments based on a written record. Other tacts which may be utilized to improve the reliability of the data include the development of descriptors to

characterize low, average, and high scores [illegible], and [illegible] patients to obtain a consistent [illegible] available [illegible] with a checklist of case material which ensures that certain behaviors can be observed (e.g. physical examination, priority ranking of problems, particular aspects of management), and the use of evaluations based on more than one case to provide a larger sample of the range of clinical behaviors.

In contrast, the long-term estimates were characterized by extremely variable reliability, primarily because the instrument possessed low discriminatory power, resulting in a narrow range of scores.

The analysis of validity substantiates these observations, in that correlations between SE and LT estimates were low, and not statistically significant. Furthermore the expected progression with educational level was present with the SE and LT estimates, but the average change with SE evaluations was 2-3 times the change observed in LT assessments.

Conclusions:

The data presented has serious implications at a time when certifying boards are recognizing formative evaluation as a component of the certification process. The assessment of clinical skills by supervising faculty, who complete a form on a periodic basis, has shown to have little value as an evaluative instrument, and if formative evaluation is to effectively provide information, it will be necessary for educators to examine in detail a number of alternate methodologies for achieving this evaluation. One alternative is the single encounter assessment, which although retaining the liabilities of subjective evaluation, appears to be more reliable than the supervisor's assessment.

It is intended to repeat this analysis in the near future, using a larger sample of about fifty residents in internal medicine, and examining the reliability of the assessment form used by the Royal College of Physicians and Surgeons of Canada. If the results of this study are verified with the larger sample, it will be the task of medical educators to develop and test alternative evaluation modalities, such as the single encounter assessment, and examine possible ways in which the reliability of these methods can be improved.

1. Wingard J.R. & Williamson J.W., Grades as Predictors of Physician's Career Performance: An Evaluative Literature Review, J. Med. Educ 48 311-322, 1973
2. Lamont C.T. & Heinnen B.K.E., The Use of Simulated Patients in a Certification Examination in Family Medicine, J. Med. Educ. 47, 789-795, 1972
3. Levine H.G. & McGuire C., Role Playing as an Evaluative Technique, J. Educ. Meas. 5, 1-7 1968
4. Assessment Form of the Royal College of Physician & Surgeons of Canada
5. "Evaluation in the Continuum of Medical Education", Report of the Committee on Goals & Priorities of the National Board of Medical Examinaers - Philadelphia, June, 1973
6. Goran M.J. et al, "The Validity of Patient Management Problems", J. Med. Educ. 48 171-177 1973
7. Norman G. R. et al, "Experience with a Single Encounter Assessment of Clinical Skills" - to be published
8. McNemar Q., "Psychological Statistics, p. 168, 1968, J. Wiley & Son, New York

TABLE I — Inter-observer reliability of S.E. and L.T. evaluation

Single Encounter			Long-Term	
	Category	Reliability	Category	Reliability
Interview Skills	1 Introduction	.67	1 History i) Interview Skills	.16
	2 Vocabulary	.47	ii) Medical Inquiry	.28
	3 Facilitation	.68	iii) Behavioral Ass.	-.006
	4 Problem Orientation	.12	2 Physical Exam	.13
History	5 Scope	.66	3 Diagnosis i) Judgment	.28
	6 Accuracy	.64	ii) Knowledge	-.09
	7 Problem Orientation	.77	4 Lab Utilization	.09
Physical	8 Scope	.71	5 Written Record	0
	9 Accuracy	.56	6 Management i) Emergency Care	.69
	10 Technique	.84	ii) Continuing Care	0
Problem	11 Specificity	.46	iii) Team Utilization	-.08
Formulation	12 Accuracy	.53	7 Technical Abilities	.43
	13 Priority Ranking	.005	8 Team Relationships	.77
	14 Investigations	.37	9 Patient Physician Relation	-.04
Management	15 Treatment	.34	10 Personal Qualities	
	16 Disposition	.08	i) Administration	-.63
Dr.-Pt.	17 Response to Pt.	.65	ii) Teaching	-.19
Interaction	18 Respect to Pt.	.47	iii) Self-learning	.04
			iv) Responsibility	.27
			v) Honesty	.18

TABLE II — Validity of S.E. and L.T. Evaluations

	Category	SE	LT	Correlation SE-LT	Change Oct.-Mar. SE	LT
A	Interview Skills	1,2,3	1i	.08	.35	.03
B	History	4,5,6	1ii,1iii	.38	.43	.15
C	Physical Exam	7,8,9,10	2	.20	.32	.09
D	Problem Formulation	11,12,13	3i,3ii	.30	.14	.08
E	Investigations	14	4	.03	.35	.23
F	Management	15,16	6i,6iii	-.08	.24	.08
G	Dr.-Pt. Relation	17,18	9	.36	.05	.40

FIGURE I — Distribution of Raw Scores